Proceedings of AMIA Annual Symposium 1998;:815-819.

Beyond Synonymy: Exploiting the UMLS Semantics in Mapping Vocabularies

Olivier Bodenreider, M.D., Ph.D.[1], Stuart J. Nelson, M.D.[2],
William T. Hole, M.D.[2], H. Florence Chang, M.S.[1]
[1]Management Systems Designers, Vienna, Virginia
[2]National Library of Medicine, Bethesda, Maryland

The Unified Medical Language System (UMLS) contains semantic information about terms from various sources, each concept can be understood and located by its relationships to other concepts: this is a result of the organizing principle of semantic locality. We describe a method in which the semantic relationships between concepts are used to map concepts from different vocabularies in the UMLS. Applied to mapping concepts to MeSH, this method is able to map 50 to 65% of the non-MeSH concepts to MeSH. A manual review of the mapping shows a relevance rate of 61%. Causes of failure include a lack of consistently represented relationships in the UMLS, and some inconsistencies in the categorization of the concepts. The limits of this method are discussed, as well as possible adaptations for other uses.

INTRODUCTION

Translating terms from one medical terminology to another is a common but non trivial problem. In the past, several methods have been proposed in which the Unified Medical Language System (UMLS)[1] is used as a source of knowledge useful to provide the translation. For example, Cimino described how to use part of this knowledge to convert ICD9-CM terms into MeSH terms.[2]

The representation of meaning in the UMLS makes it possible for users to explore the semantic neighborhood of one concept in order to reach the nearest neighbor in one given source. The different expressions of semantic links between concepts represent one of the organizing principles of the UMLS: semantic locality. These dimensions of semantic locality include term information (synonymy, hypernymy, hyponymy), contextual information in a particular source, co-occurence of terms in the medical literature, and the categorization of the concepts in a semantic network.[3, 4]

This paper examines the use of three of the dimensions of semantic locality in the UMLS in order to find the MeSH terms most closely related to any given UMLS concept. This work is part of the Indexing Initiative, an ongoing effort of the National Library of Medicine to investigate automated indexing methods as a partial or complete substitute for current indexing practices.

BACKGROUND

The UMLS is intended to help health professionals and researchers use biomedical information from different sources.[5] While the structure of each source vocabulary is preserved, terms which are equivalent in meaning are clustered into a unique concept. Furthermore, interconcept relationships, either inherited from the source vocabularies or specifically generated, give to the UMLS Metathesaurus additional semantic structure. This structure can be visualized as a graph in which concepts are the nodes and interconcept relationships the links between nodes. The UMLS Semantic Network is a network of semantic types used to categorize each concept. The relationships between semantic types within the Semantic Network describe the possible relationships between the concepts categorized by these semantic types within the Metathesaurus.[4] Although the UMLS is mostly a collection of precoordinated terms of various granularity, the associated expressions (ATXs) created by indexers provide a translation of some complex concepts to expressions in other vocabularies, using elementary concepts combined with both logical operators and possibly, in mappings to MeSH (Figure 1), main heading (MH) and subheading (SH) combinations. Synonymy and lexical matching techniques are used to link terms together. At the concept level, beyond synonymy, the semantics of the UMLS can be exploited in mapping vocabularies.

The 1998 version of the UMLS contains 476,313 concepts from more than 30 vocabularies.[1] Of these, 185,406 concepts are to some extent considered MeSH terms (they contain 'MSH98' in the SAB field of the MRSO file). While only 19,000 of them are MeSH main headings, the others are entry terms, qualifiers, or come from the large list of supplementary chemical terms.[6] 7,073 concepts are described by at least one associated expression, most of them coming from MeSH. Table 1 shows the distribution of interconcept relationships. About 2%

of the concepts do not have any relationship with another concept.

Table 1 - Distribution of interconcept relationships in the UMLS: number of concepts having such a relationship.

Type of relationship		number
parents	PAR	273,551
children	CHD	52,533
siblings	SIB	153,121
broader concepts	RB	227,074
narrower concepts	RN	27,990
similar concepts	RL	155,548
allowable qualifiers	AQ	18,297
qualified by	QB	98
unlabeled relationship	RO	89,869

METHODS

Three basic approaches can be used to map a UMLS term to MeSH: through synonyms, through associated expressions, and through interconcept relationships. These approaches can be combined into a strategy that maximizes both specificity (selected MeSH terms are relevant) and sensitivity (the number of concepts that fail to be mapped to MeSH is small).

Strategy

The overall strategy can be understood as involving four steps. For a given UMLS concept, referred to as the source concept (SC), the path to the most closely related MeSH terms utilizes the following steps in this order:

1. A MeSH term is a synonym of the SC. The two terms share the same identifier in the Metathesaurus (CUI). This MeSH term is selected and no further search is performed.
2. An associated expression (ATX) provides a translation of the SC. The ATX can be understood as an expression tree in which leaves are elementary concepts and nodes logical operators or main heading to subheading relationship indicators (Figure 1). For mapping to MeSH headings, all MeSH leaves are selected, except those under a negative (NOT) operator. For example, the concept "Mumps pancreatitis" is mapped to the following MeSH terms: "Mumps" and "Pancreatitis" (main headings), "complication" and "etiology" (subheadings).
3. The SC has hierarchically related concepts from which MeSH terms can be selected. This method is detailed under mapping algorithm.
4. Finally, if no MeSH term can be found from the ancestors, the non-hierarchically related concepts (RO concepts) are explored. These concepts are related to the SC, but the exact nature of this relationship has not been explicitly given. Steps 1 to 3 are then applied to each RO concept linked to the SC. For example, "Choroidal detachment, NOS" is related to the MeSH term "Retinal Detachment".

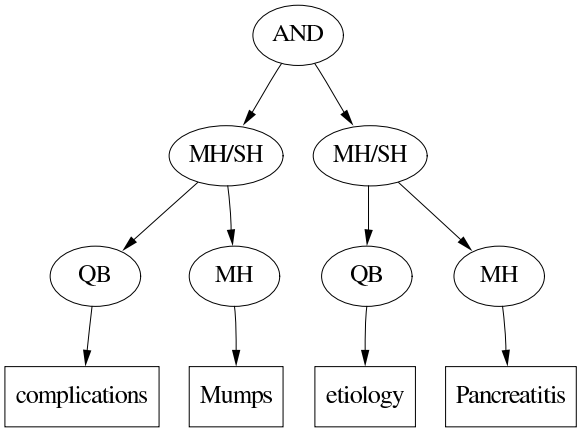

Figure 1 - Expression tree for the associated expression describing the concept: "Mumps pancreatitis". The main heading (MH) is qualified by (QB) a subheading (SH). The 2 MH/SH expressions are combined with a logical operator (AND).

Mapping algorithm

The mapping algorithm can be visualized as building a graph of ancestors, using the SC as the initial point, or seed, in building this graph. Then from this graph the closest MeSH terms are selected. Other concepts than the SC itself can be used to start populating the graph of ancestors. Children and narrower concepts of the SC can be used together as the seed of the graph when no MeSH terms can be found from the graph seeded by the SC. Failing to find a MeSH term by that method, a new graph is generated, using siblings of the SC.

In the event of using concepts other than the SC itself as the seed for the graph, the concepts chosen as the seed must be compatible in semantic type assignment. Compatibility is defined as the situation where at least one of the semantic types (STs) of the concept is identical to or has an "inverse_isa" relationship in the Semantic Network to at least one of the STs of SC. Siblings of the SC must have at least one ST in common to be used as seed of the graph.

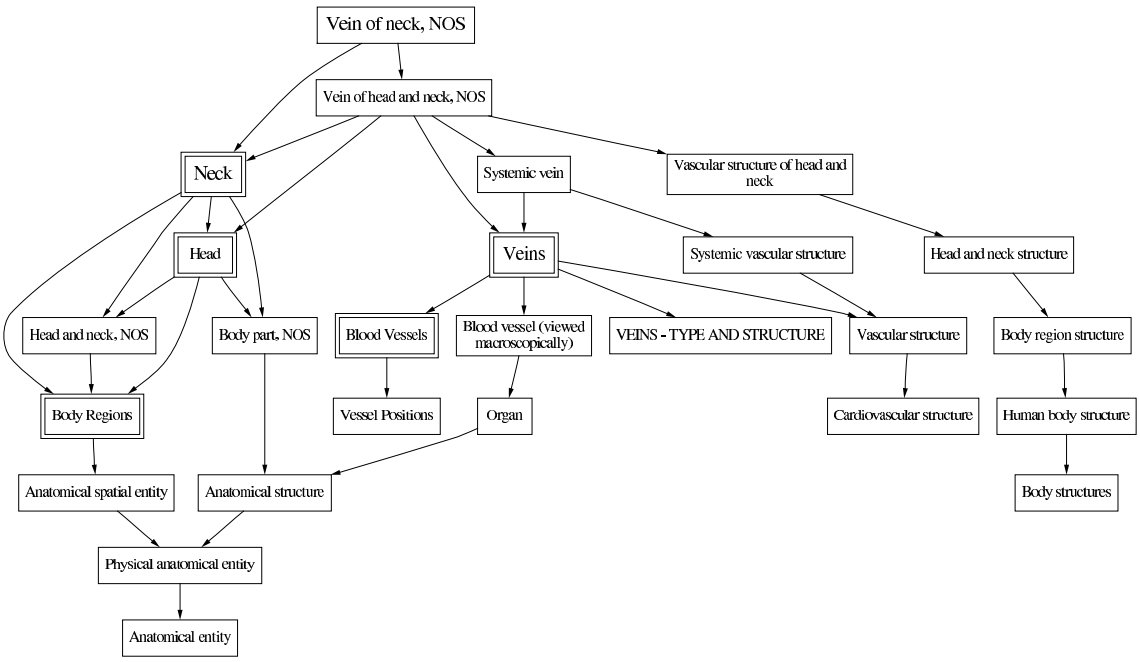

Figure 2 - Graph of the ancestors of "Vein of neck, NOS". MeSH terms are double framed. The selected MeSH terms are "Neck" and "Veins". Arrows point to parents or broader concepts.

Step 1: Building the graph of the ancestors of the SC. The ancestors of a given concept can be represented as a directed graph, ideally acyclic. Starting from the seed, its parents and broader concepts are added to the graph. Then, recursively, parents and broader concepts of all newly added concepts are added, until no new concept can be found.
To prevent non relevant concepts from being added to the graph, the semantic types of any concept added to the graph must be compatible with those of its direct descendant in the graph.

Step 2: Selecting MeSH terms from the ancestors. The graph of the ancestors is first restricted to MeSH terms (synonyms or from associated expressions). Then, to prevent MeSH terms to come only from one part of the seed, that is one particular child or sibling, selected MeSH terms must be common to a certain percentage of the seed concepts that have MeSH ancestors. Finally, MeSH candidates that are ancestors from each other are removed. The selected MeSH terms are thus insured to be semantically as close as possible to the SC. This measure of closeness relies on semantics rather than on the number of nodes between the two concepts, which is biased by the difference in granularity between components of the UMLS.
Figure 2 shows how "Neck" and "Veins" are selected from the ancestors of "Vein of neck, NOS". Although the MeSH term "Head" is at the same distance as "Veins", it is not selected because it is an ancestor of another selected term ("Neck").

This algorithm was implemented in Perl. UMLS data extracted from the relational tables (MRCON, MRREL, MRATX, MRSO, MRSTY, SRSTRE1) were put in B-Tree files.

Testing
2 sets of UMLS concepts were selected to test the algorithm. Set 1 consisted of 1,000 randomly selected concepts from the 290,907 concepts unrelated to MeSH. Set 2 consisted of 1,036 unique concepts unrelated to MeSH extracted from a random selection of 200 citations (title and abstract) from MEDLINE. The text was mapped to the UMLS using the MetaMap program.[7] 66% of the original concepts were mapped directly to MeSH.
The quality of the mapping was evaluated by a manual review. The following classification was used to describe the quality of the mapping: "relevant" means that the selected MeSH terms were relevant to the source concept, even if a more specific term was available; "non relevant" means that none of the selected MeSH terms was a correct map; "more or less relevant" means that the selected MeSH terms were not irrelevant to the source concept, but were either far ancestors or only part of

the ancestors needed to correctly describe the source concept.

RESULTS

The results of the mapping of 2 sets of UMLS concepts are summarized in Table 2.

Every mapping and failure from Set 2 was reviewed. 61% of the MeSH terms coming from the ancestors were relevant. In about 28% of cases, the selected MeSH term was very broad and did not give more information on the source concept than its semantic type could be expected to (e.g. "Serotonin measurement" mapped to "Laboratory Procedures", entry term for "Laboratory Techniques and Procedures"). 11% of the mappings were not relevant. MeSH terms coming from associated expressions are essentially always relevant.

Table 2 - Approaches used for the mapping: total number of concepts mapped, whatever the relevance of the mapping.

Type of mapping	Set 1	Set 2
associated expressions	26	20
ancestors, from parents	530	204
ancestors, from children	18	34
ancestors, from siblings	8	22
from RO concepts	69	231
failure	349	525
total	1,000	1,036

DISCUSSION

Not surprisingly, the most part of non relevant MeSH terms came from the use of RO concepts (step 4 of the strategy). However, about two out of three MeSH terms from this path were relevant. Half of the source concepts reaching this path were isolated adjectives, simply linked to their corresponding MeSH nominal equivalent.

It appears that non relevant MeSH terms often come from a unique path in the graph built from children or siblings. This can occur when only one of the concepts from the seed of the graph has MeSH ancestors. In this case, only one part of the semantics is emphasized, which is not necessarily common to the source concept. For example, "Holt-Oram syndrome" (a multiple malformation syndrome with limb defect) is wrongly mapped to "Facial Paralysis", because "Mobius Syndrome" (a multiple malformation syndrome with facial paralysis) is the only sibling from which a MeSH term can be reached.

Failure to map to MeSH or mapping to broad MeSH terms are usually caused by a lack of relationships being represented in the UMLS. Methods relying on semantic principles assume that all the relationships are expressed. A certain lack of relationships has already been identified in other studies.[8, 9]

Some of the relationships assigned between concepts in a source vocabulary are not always consistent. Two concepts close in meaning may not be mapped to the same MeSH terms. For example, the two following concepts from SNOMED International "Thrombectomy with catheter of celiac artery by abdominal incision" (TCA) and "Thrombectomy with catheter of popliteal-tibio-peroneal artery by leg incision" (TPA) are not mapped consistently to MeSH. TCA is correctly mapped to "Thrombectomy", while TPA is mapped to "Operative Procedures" (entry term for "Surgical Procedures, Operative"). Although "Thrombectomy" (also found in SNOMED) is not related to these concepts in SNOMED, TCA was linked to "Thrombectomy" as a narrower concept during the UMLS building process, so that "Thrombectomy" could be selected as one of its ancestors. However, the detection of related concepts is not always performed consistently, and TPA, which has no explicit relationship to "Thrombectomy", can not be mapped to the same MeSH term as TCA.

As noted in other studies, inconsistencies in the categorization of the concepts are a source of failures in methods relying on semantics.[9, 10] Since the compatibility of semantic types between related concepts is checked in the graph of ancestors, a lack of semantic types or, more often, inconsistencies in the categorization of the concepts can result in the inappropriate rejection of some valuable parents from the graph.

Assuming that the mapping method is reliable enough, the comparison of the mapping to MeSH of several concepts known to be close in meaning could be a way to detect inconsistencies in the categorization of concepts in the UMLS.

This algorithm can be tuned from a strict mode (only relevant MeSH terms, but with large number of failures) to a relaxed mode (all possible MeSH terms, some of them being not relevant). The method that we described is a medium mode, suitable to our goal: the selected MeSH terms are intended to be filtered and clustered according to additional information such as the frequency of each term in the source text and how often these terms co-occur in the medical literature.

Specificity can be increased by allowing the graph of ancestors to be built only from the source concept itself and not from its children or siblings. Not using the RO concepts can also help limit the percentage of non relevant MeSH terms. In addition, taking into

account relationship attributes (type of relationship, e.g. "isa") could help keep only those hierarchical relationships which are meaningful to our purpose. For example, while both "Aortic Arch" and "Arteries" are parents of "Brachiocephalic Trunk", we would like to retain "Arteries", which describes a class ("isa" relationship) and not "Aortic Arch" ("branch_of" relationship), which only illustrates the anatomical relationships between two instances of the Arteries class. Unfortunately, less than 5% of interconcept relationships are explicitly described by attributes such as "isa", "branch_of", etc.

In the other hand, *sensitivity* can be increased by using a broader notion of the semantic type compatibility, or by not checking the semantic types of the ancestors. Similar concepts (candidates for synonymy but not currently reviewed) can also be used in addition to the RO concepts.

CONCLUSION

By focusing on the other principles of semantic locality, rather than solely on synonymy, we observed that:

- It is helpful to think of the UMLS as a graph whose nodes (concepts) have multiple facets (terms) rather than as a collection of terms clustered into concepts.
- The tools needed to manipulate the knowledge are closer to graph manipulation tools than to programs implementing lexical knowledge.
- Hypertext-based tools or browsers make it possible to explore the knowledge by navigating,[11] while terminology servers continue to help users find and express one particular piece of knowledge.

Semantically driven methods are suitable to map vocabularies in the UMLS. However, these methods require an ideal UMLS to have both maximal sensitivity and specificity. This ideal UMLS would identify and label any possible interconcept relationship. Meanwhile, accepting some limits, the current UMLS with its annual enhancements already gives useful results.

Acknowledgments

All graphs have been drawn using the Graphviz package made available by AT&T Corp.

This work was supported in part by 3M Laboratories, CHU and Faculté de Médecine de Nancy (France), and by the following associations: Association des Utilisateurs des Nomenclatures Nationales et Internationales de Santé (AUNIS), Collège des Praticiens Spécialistes en Information et Communication Médicales (COPSICOM), Information Médicale et Gestion des Établissements (Groupe IMAGE) and Contribuer à la Recherche et à l'Innovation au Service du Traitement et de l'Analyse des Langages utilisés dans les Systèmes de Soins (CRISTAL'S).

References

1. UMLS Knowledge Sources. (9th ed.) Bethesda (MD): National Library of Medicine, 1998.
2. Cimino J, Johnson S, Peng P, Aguirre A. From ICD9-CM to MeSH using the UMLS: a how-to guide. Proc Annu Symp Comput Appl Med Care 1993:730-4.
3. Nelson S, Tuttle M, Cole W, et al. From meaning to term: semantic locality in the UMLS Metathesaurus. Proc Annu Symp Comput Appl Med Care 1991:209-13.
4. McCray A, Nelson S. The representation of meaning in the UMLS. Methods Inf Med 1995;34(1-2):193-201.
5. Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. Methods Inf Med 1993;32(4):281-91.
6. Medical Subject Headings. Bethesda (MD): National Library of Medicine, 1998.
7. Aronson A, Rindflesch T, Browne A. Exploiting a large thesaurus for information retrieval. Proceedings of RIAO 94, 1994:197-216.
8. Burgun A, Botti G, Bodenreider O, et al. Methodology for using the UMLS as a background knowledge for the description of surgical procedures. Int J Biomed Comput 1996;43(3):189-202.
9. Cimino J. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51.
10. Bodenreider O, Burgun A, Botti G, Fieschi M, Le Beux P, Kohler F. Evaluation of the Unified Medical Language System as a medical knowledge source. J Am Med Inform Assoc 1998;5(1):76-87.
11. Tuttle M, Cole W, Sheretz D, Nelson S. Navigating to knowledge. Methods Inf Med 1995;34(1-2):214-31.